EmoDB Speech Classification

First Model

Format: As we deal with wav files and need to tackle a speech recognition challenge, mfcc format

Model: the first architecture was inspired by this repo who already achieved a validation accuracy of 95% on the same dataset, limited to 4 classes (Angry,Happy,Neutral,Sad). It consists of a 4-layers CNN treating the mfcc data as a single-channeled image.

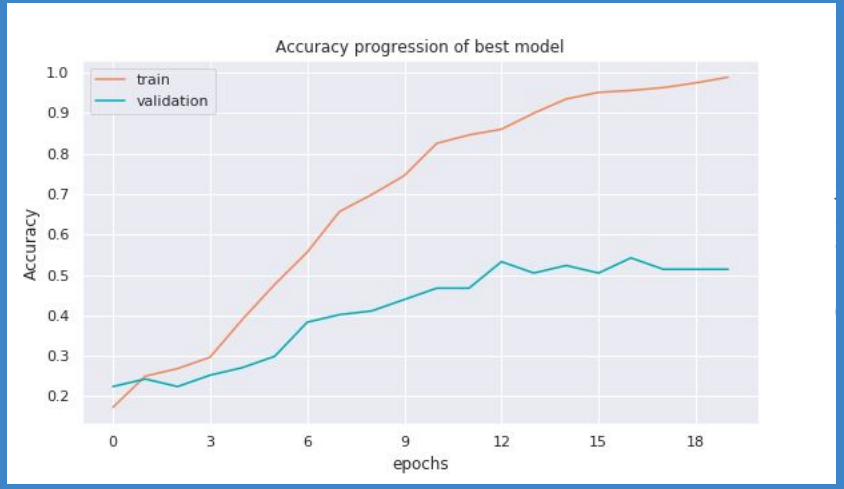
Hyper-parameters: the hyper parameters were tuned by GridSearch using the skorch library

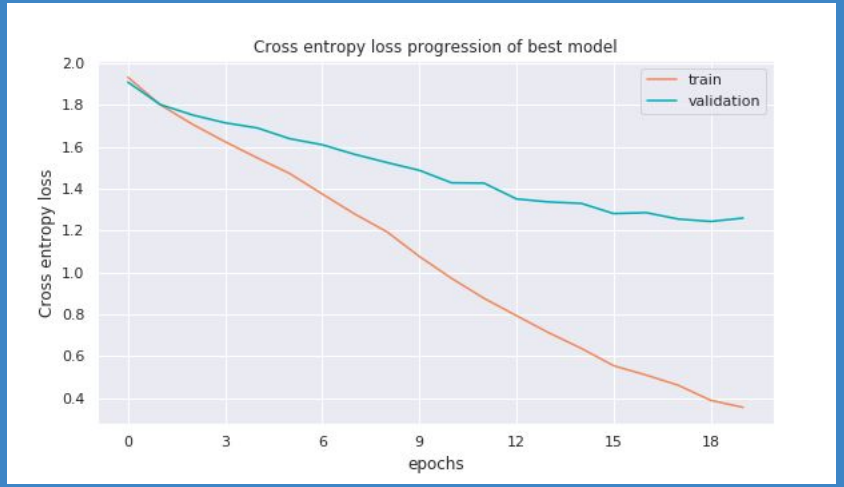
First Result

As we can see, the validation accuracy reaches a ceiling around 50%, indicating is an overfitting issue. The network seems to have enough representational power but need some type of regularization/data augmentation.

In order to tackle this issue, 3 feature augmentation techniques were implemented:

- Temporal Shift:
- Gaussian Noise addition:
- Pitch tuning



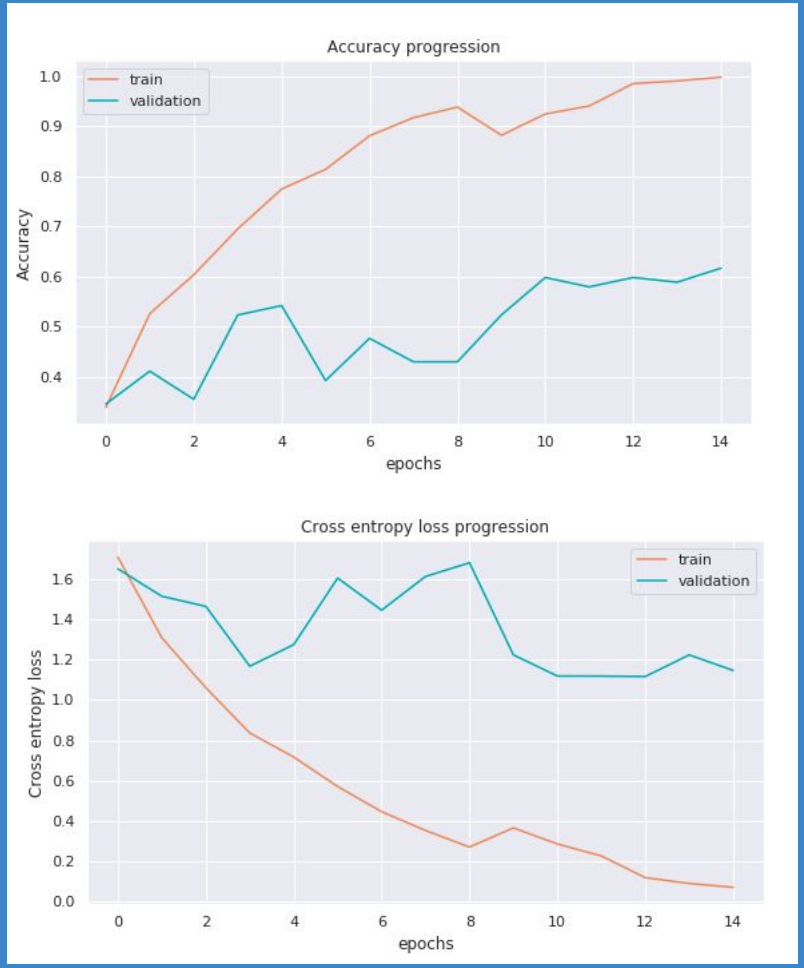


Regularized Model

The regularization seems to have allowed the validation to increase of ~10%

Final Model:

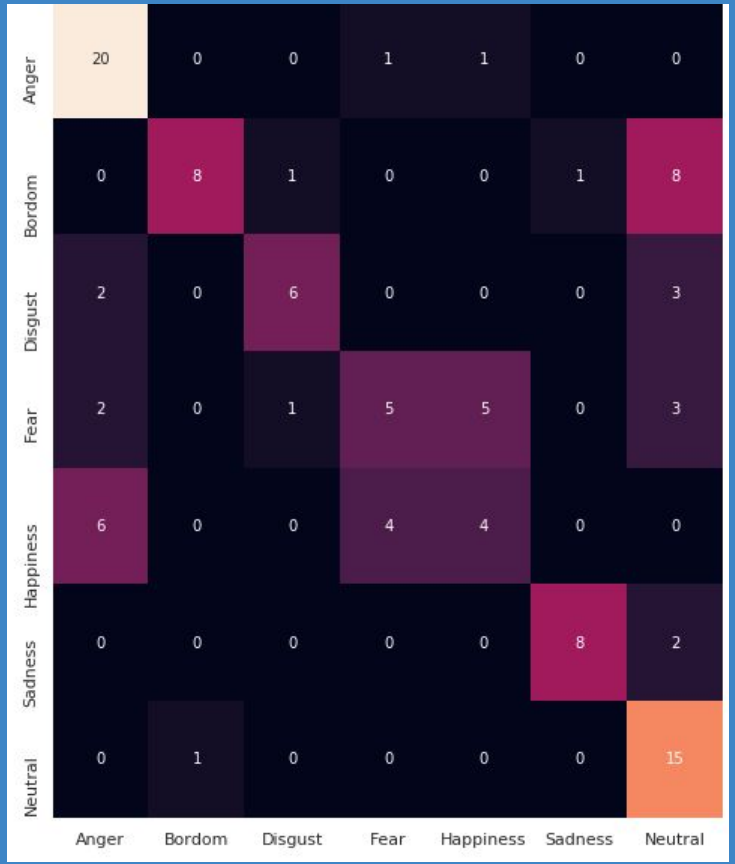
- train accuracy: 98% +- 2.3 (4Fold CV)
- val accuracy: 58% +- 1.4 (4Fold CV)



Misclassification Analysis

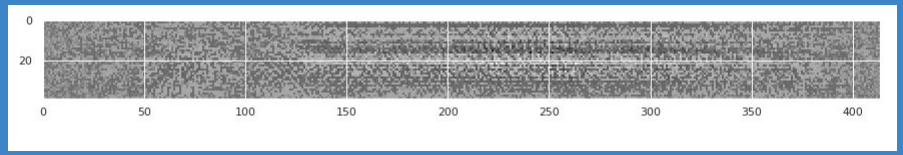
It seems like the main responsible features for bad validation accuracies are Bordom interpreted as Neutral and Happiness as Anger.

This does make sense, as the 2 formers emotions are pretty quietly pronounced whereas the latter are more energetic, suggesting that the network potentially focuses on the exclamation rate to classify the samples.



Max Activation Layer

In order to gain insights on the type of features the network focuses on to do its prediction, we use a gradient descent algorithm on the pixels of the input image to maximize the ouput of the various layer. The goal is to recover a mfcc which would show a particular pattern.



Unfortunately, due to relatively poor validation accuracy, the generated image does not reveal any significant feature. However this could happen with better hyper-parameter tuning.